Economics and Philosophy, **28** (2012) 43–71
doi:10.1017/S0266267112000119

SOCIAL JUSTICE, GENOMIC JUSTICE AND THE VEIL OF IGNORANCE: HARSANYI MEETS MENDEL

SAMIR OKASHA
University of Bristol, UK
samir.okasha@bristol.ac.uk

John Harsanyi and John Rawls both used the veil of ignorance thought experiment to study the problem of choosing between alternative social arrangements. With his 'impartial observer theorem', Harsanyi tried to show that the veil of ignorance argument leads inevitably to utilitarianism, an argument criticized by Sen, Weymark and others. A quite different use of the veil-of-ignorance concept is found in evolutionary biology. In the cell-division process called meiosis, in which sexually reproducing organisms produce gametes, the chromosome number is halved; when meiosis is fair, each gene has only a fifty percent chance of making it into any gamete. This creates a Mendelian veil of ignorance, which has the effect of aligning the interests of all the genes in an organism. This paper shows how Harsanyi's version of the veil-of-ignorance argument can shed light on Mendelian genetics. There turns out to be an intriguing biological analogue of the impartial observer theorem that is immune from the Sen/Weymark objections to Harsanyi's original.

1. INTRODUCTION

In *A Theory of Justice*, Rawls (1971) invoked the device of an 'original position' to study the problem of social justice. He imagined someone

Acknowledgements: Thanks to Ken Binmore, Cedric Paternotte, Jonathan Grose, Ellen Clarke and Johannes Martens for discussion. Thanks to an anonymous referee for *Economics and Philosophy*, and to the editor, for detailed comments. Thanks also to audiences at Oxford, Vienna and Bristol where versions of this material were presented. Financial support from the UK Arts and Humanities Research Council and from the European Research Council is gratefully acknowledged.

forced to choose between alternative social arrangements from behind a 'veil of ignorance', i.e. without knowing which member of society she will become. Rawls maintained, controversially, that the rational agent would choose the social alternative which maximized the prospects of the least well-off member of society - the maximin principle.

As is well known, Rawls was not the first to use a veil of ignorance argument to think about social justice. In two famous papers, Harsanyi (1953, 1955) imagined a 'sympathetic, impartial observer', again tasked with choosing between social alternatives from behind a veil; a still earlier version of the idea was sketched by Vickrey (1945). By contrast with Rawls, Harsanyi argued that the veil-of-ignorance thought experiment leads inevitably to utilitarianism, an argument that has come to be called the 'impartial observer theorem'. Harsanyi arrived at this conclusion by assuming that the impartial observer has an equal chance of becoming any individual, and chooses between social alternatives in accordance with expected utility maximization.

Most recent commentators regard Harsanyi's treatment of the veil of ignorance as superior to Rawls, given Rawls's largely unmotivated rejection of orthodox decision theory. But even so, controversy surrounds the impartial observer theorem. A number of authors, notably Sen (1976, 1977, 1986), Weymark (1991), Roemer (1998) and Mongin (2001), have argued that Harsanyi's theorem does not constitute a good argument for classical utilitarianism, for reasons explained below. There is a large and well-known literature on this issue.

What is much less well-known is that the veil-of-ignorance concept has also surfaced in evolutionary biology, in a context quite remote from the traditional Rawls/Harsanyi debate. During meiosis – the cell-division process by which sexually reproducing organisms make gametes – the chromosome number is halved: only one of each chromosome pair is passed to each gamete. Most of the time meiosis is 'fair', so that any particular gene has a 50% chance of making it into any gamete – a fact known as Mendel's law of segregation. The law has profound evolutionary consequences, as it equalizes the interests of all the genes in the organism, ensuring they work for the common good; see section 5. Indeed fair meiosis, or Mendelian segregation, is arguably a prerequisite for the very existence of cohesive organisms such as ourselves.

That fair meiosis serves to equalize genes' interests was first emphasized by Leigh (1971), and is now an accepted biological principle. But it is only recently that the intriguing analogy between fair meiosis and the veil of ignorance has come to light. In both, randomization is used to deprive self-interested agents (genes and individuals) of information about their identity, forcing them to adopt an impartial perspective. This analogy is noted in passing by the biologists Haig and Bergstrom (1995) and Frank (2003), who cite both Rawls and Harsanyi, but the only

extended discussion of which I am aware is in a semi-popular book by Mark Ridley (2000), who mentions only Rawls.[1]

The aim of this paper is to consider in detail whether the traditional veil-of-ignorance concept can be used to understand Mendelian genetics. Can the Harsanyi/Rawls thought experiment shed light on the actual process of fair meiosis? I suggest that it can, and in particular that there is an interesting biological analogue of the impartial observer theorem. This may sound like a rather unlikely project, so it is worth briefly explaining the motivation for it, which is three-fold.

Firstly, there are many interesting parallels between economics and evolutionary biology, arising because the concept of *utility* in the former plays a similar role to the concept of *fitness* in the latter. The importing of game-theoretic concepts into biology is the best known illustration of this role-isomorphism, but is not the only one. The current project also trades on the utility/fitness connection, and is in part an attempt to better understand the elusive relation between these two concepts.

Secondly, there is plenty still to say about how the veil-of-ignorance idea applies to genetics. Ridley's discussion, though illuminating, is compromised by his considering only Rawls's version of the veil of ignorance. Since Harsanyi arguably had the more coherent version, his is the better place to look for a link with Mendelian genetics. Additionally, Ridley's treatment contains a subtle confusion between proximate and ultimate explanations, as I argue below, and ends up locating the link in the wrong place.

Thirdly, my project fits naturally with a philosophically exciting way of thinking about Darwinian evolution, which treats genes as if they were rational agents trying to maximize a utility function. This approach, aptly dubbed the 'heuristic of personification' by Sober (1998), permeates modern evolutionary thinking and is of undoubted value for some purposes. But doubts over its legitimacy have often been raised.[2] By modelling genes as individuals behind a veil of ignorance, I hope to illustrate both the power and the limitations of the personification heuristic.

The structure of this paper is as follows. Section 2 expounds Harsanyi's version of the veil-of-ignorance argument. Section 3 outlines the now-standard objections to the argument due to Sen (1986) and

[1] Skyrms (1996) discusses what he calls the 'Darwinian veil of ignorance' in his work on the evolution of the social contract, but his use of this notion is unrelated to the one explored here, as it has nothing in particular to do with Mendelian genetics. The same is true of Binmore's (2006) attempt to locate the Rawls/Harsanyi 'original position' in an evolutionary context.

[2] Personifying genes as an aid to evolutionary reasoning was employed extensively by Dawkins in *The Selfish Gene* (1976), though the roots of the idea are found in W.D. Hamilton's papers from the 1960s (1964). Haig (1997) presents a sophisticated defence of the heuristic, and Godfrey-Smith (2009) a sophisticated critique.

Weymark (1991). Section 4 provides some biological background on Mendelian genetics. Section 5 considers the evolutionary function of fair meiosis, and its implications. Section 6 explores two different ways of modelling fair meiosis in Harsanyi's framework. Section 7 examines parallels, conceptual and formal, between Harsanyi's impartial observer argument and our evolutionary analogue. It is argued that the Sen/Weymark critique of Harsanyi's argument does not apply to the evolutionary analogue. Section 8 compares my analysis with Ridley's. Section 9 concludes.

2. HARSANYI'S IMPARTIAL OBSERVER ARGUMENT

Harsanyi's own formulation of the impartial observer argument is rather elliptical, so I draw on the careful reconstructions by Weymark (1991) and Mongin (2001). The context is a standard social choice setting. There is a finite set of social alternatives S, which could for example be alternative distributions of resources among society's members. There is a finite set of individuals I. Each individual has a (weak) preference order over the alternatives in S. The preference order of the i^{th} individual will be denoted R_i; so '$x R_i y$' means that the i^{th} individual weakly prefers social alternative x to y.

Harsanyi assumes further that each individual has a preference order over the set of lotteries whose prizes are the members of S, i.e. the set of probability distributions over S, denoted ΔS. These lotteries will be called simple lotteries (to be contrasted with the extended lotteries below). Thus for example if $S = \{x, y, z\}$ then one simple lottery is $< P(x) = \frac{1}{4}, P(y) = \frac{1}{2}, P(z) = \frac{1}{4} >$. Obviously, any alternative x in S can be identified with the simple lottery that gives x probability 1 of occurring. The point of considering preferences over lotteries is to license the introduction of cardinal utility. Each individual's preferences over ΔS are assumed to satisfy the von Neumann–Morgenstern (vNM) axioms, hence can be represented by an expectational utility function unique up to affine transformation.

Harsanyi then imagines a hypothetical 'impartial observer' who is sympathetic to the interests of the members of society, and is able to imagine himself in the role of any member. The observer can evaluate 'extended alternatives' of the form 'being individual 2 in social alternative x'. An extended alternative is thus an ordered pair of an individual and a social alternative; the set of all extended alternatives is $I \times S$. The observer is assumed to have a preference order over this set; thus he can make judgements such as 'I would prefer to be individual 2 in alternative x than individual 3 in alternative y'.

Harsanyi next considers the set of lotteries over the extended alternatives, or the 'extended lotteries'[3], denoted $\Delta(I \times S)$. He assumes that the

[3] This terminology comes from Weymark (1991).

impartial observer has a preference order over the extended lotteries that satisfies the expected utility axioms, so is representable by a vNM utility function. We let R_o denote the observer's preference order over $\Delta(I \times S)$, and let u_o denote a particular vNM utility representation of R_o.

One subset of the extended lotteries is of particular significance: the impartial extended lotteries in which the observer has an equal chance of becoming any member of society.[4] Thus if there are n individuals in set I, the lottery $< P(x, i) = \frac{1}{n}$ for all $i >$ is the impartial extended lottery in which social alternative x definitely occurs and the observer has an equal chance of becoming any individual. (The corresponding simple lottery is $< P(x) = 1 >$, i.e. alternative x for certain.) Harsanyi assumes that the probability that the observer becomes any given individual is independent of the probability that any particular social alternative occurs. This means that for every simple lottery there is a unique impartial extended lottery that corresponds to it, i.e. which yields the same marginal probabilities for the social alternatives as the simple lottery.

A simple example may help illustrate. Suppose there are two social alternatives x and y, and two individuals. So $S = \{x, y\}$, and $I = \{1, 2\}$. There are four extended alternatives: $\{(1, x), (1, y), (2, x), (2, y)\}$. Consider the simple lottery L_1 in which x occurs with probability $2/3$ and y with $1/3$, i.e. $L_1 = < P(x) = \frac{2}{3}, P(y) = \frac{1}{3} >$. The impartial extended lottery E_1 that corresponds to L_1 is then: $< P(1, x) = \frac{1}{3}, P(1, y) = \frac{1}{6}, P(2, x) = \frac{1}{3}, P(2, y) = \frac{1}{6} >$. Note that E_1 gives the observer equal chances of becoming either person, and gives alternatives x and y the same probability of occurrence as does L_1. Also, E_1 makes the probability that the observer is a given person independent of the probability that a given social alternative occurs.

Since the impartial observer is sympathetic to the interests of society's members, Harsanyi proposes a link between the preferences of the observer when he is imagining himself to be a given individual, and the preferences of that individual himself. The observer's ordering of extended lotteries in which he is definitely individual i should coincide with individual i's ordering of the corresponding simple lotteries. This is Harsanyi's *principle of acceptance*. The underlying idea is that each individual's personal preferences are sovereign, so when the observer imagines himself in the shoes of a given individual he thereby imagines himself to have that individual's personal preferences.

From the observer's preference ordering over the impartial extended lotteries, Harsanyi then derives a social preference over the simple lotteries, and thus over the social alternatives themselves. He simply

[4] Giving the observer an equal chance of becoming any individual is Harsanyi's way of modelling the observer's epistemic position from behind the veil of ignorance. Harsanyi is invoking the classical principle of indifference at this juncture, as Mongin (2001) notes.

postulates that society's preference between simple lotteries derives from the observer's preference between the corresponding impartial extended lotteries. Consider two simple lotteries L_1 and L_2 and the corresponding impartial lotteries E_1 and E_2. Harsanyi says that L_1 is socially preferable to L_2 if and only if the impartial observer would prefer E_1 to E_2. Since any social alternative is itself a (degenerate) simple lottery, this yields a way of ordering the social alternatives.

Clearly, Harsanyi's extraction of a social preference from the observer's extended preferences rests on an ethical judgement. Harsanyi thinks that society should choose between alternatives (or lotteries) according to how a sympathetic observer, with an equal chance of becoming any individual, would choose between them. Intuitively this captures our concept of social justice quite well, but its ethical standing could obviously be questioned.

The ingredients are now in place for Harsanyi to derive his utilitarian conclusion. Suppose R is the social preference order on the simple lotteries, defined via the observer's extended preference order R_o. Since R_o satisfies the vNM axioms, so does R. Thus R can be represented by a vNM utility function u, which we may call the 'social utility function'. Harsanyi then shows that the social utility function can be expressed as the average of all the individual's utility functions. So society follows an average utilitarian rule: it ranks simple lotteries (and thus alternatives) according to the average utility that they bring to members of society. This is the impartial observer theorem.

More precisely, what Harsanyi shows is this. There exist individual vNM utility functions $u_1, ..., u_n$, one for each member of society, which represent the individuals' preference orders $R_1, \ldots, R_n$; and the social utility function u which is the average of these n individual utility functions, i.e. $u(x) = \frac{1}{n} \sum u_i(x)$ for all $x \in S$, represents the social preference order R, when R is defined via the observer's extended preference R_o on the corresponding impartial lotteries.[5]

How do we find the particular individual utility functions for which this is true? They are implicitly defined by the particular choice of vNM utility function u_o to represent the observer's extended preference ordering. Suppose that $y \in \Delta(I \times S)$ is an extended lottery in which the observer is individual i for certain. Let x be the corresponding simple lottery. Then set $u_i(x) = u_o(y)$, i.e. individual i's utility for the simple lottery x equals the observer's utility for the extended lottery y.[6] Relative to the individual utility functions $u_1, ..., u_n$ defined this way, the social utility function u is given by the average utilitarian rule.

[5] This is a verbal paraphrase of theorem 9 in Weymark (1991).

[6] Since the principle of acceptance is satisfied, the function u_i defined by this procedure represents individual i's preference order.

To summarize, Harsanyi's impartial observer theorem tries to derive a utilitarian conclusion from three ingredients. These are: (i) the assumption that all individuals, including the observer, are expected utility maximizers; (ii) the principle of acceptance; and (iii) the ethical postulate that society's choice should be determined by what the impartial observer would choose from behind the veil. The difference between Harsanyi and Rawls stems chiefly from the latter's rejection of (i).

3. THE SEN/WEYMARK CRITIQUE OF HARSANYI

Though formally straightforward, the philosophical significance of the impartial observer theorem is a matter of ongoing controversy; see Mongin (2001) for a recent assessment. Sen (1986) argued that Harsanyi's theorem does not in fact establish a utilitarian conclusion, in the ordinary sense of utilitarianism, but is merely a representation theorem for social preferences. This criticism was endorsed and further developed by Weymark (1991).

The essence of the Sen/Weymark critique is that there is an irreconcilable tension in Harsanyi's argument. On the one hand he needs to assume that utility is fully interpersonally comparable, i.e. both utility levels and differences can be compared across individuals. (Difference comparability is needed for utilitarianism to be a well-defined doctrine, and level comparability is needed for the impartial observer to be able to make his judgements of extended preference.) On the other hand he assumes that 'utility' means utility in the sense of von Neumann and Morgenstern, i.e. a numerical representation of a preference relation over lotteries. Preferences are primary, on the standard vNM picture, and utility derived.

This is inherently problematic. One problem is that the vNM theory does not itself tell us how to make the required interpersonal comparisons; but there is a deeper problem. As Weymark stresses, if utility is merely a representation of preference, there is no particular reason to restrict attention to vNM utility functions (i.e. ones that are linear in the probabilities, or expectational). The vNM theorem tells us that if an individual's preference relation over lotteries satisfies certain axioms, then it *can* be represented by an expectational utility function, but it can equally well be represented by many other utility functions which are not expectational.[7] However it is essential to Harsanyi's theorem that only vNM utility functions are used to represent preferences; without this,

[7] If u is a vNM utility representation of a given preference relation R over lotteries, and $f(u)$ is any increasing transformation of u, affine or not, then $f(u)$ will also represent R. On the orthodox view, the reason for employing vNM utility functions, rather than any other, is simply mathematical convenience; as Arrow (1951) said, their merit is 'stating the laws of rational behaviour in a particularly convenient way' (p. 10).

the utilitarian conclusion does not go through. But Harsanyi offers no justification for only considering vNM utility functions; he appears to think, wrongly, that their use is mandated rather than merely permitted by the von Neumann–Morgenstern expected utility theorem.

Weymark concludes from this that for Harsanyi's argument to succeed, it must operate with a non-preference based concept of utility. For example, utility could be construed as some kind of mental or hedonic state, à la classical utilitarianism, whose relationship to preferences would then be an empirical matter. (Indeed Harsanyi at times seems to have such a concept in mind.) If such a notion of utility is granted, and certain assumptions about it are made, then Harsanyi's argument may be salvageable. The assumptions are that utility is real-valued, cardinally measurable, fully interpersonally comparable, and expectational, i.e. the utility of a lottery is its expected utility.

This is the basis for a version of Harsanyi's theorem proposed by Weymark (1991) (his theorem 10). To avoid confusion Weymark talks about well-being rather than utility, and makes the above assumptions about well-being. As in Harsanyi's original, there is a set of social alternatives, and a set of lotteries over them. Each individual has a well-being function over the lotteries. The impartial observer has a well-being function over the extended alternatives, and the extended lotteries. The 'principle of welfare identity' says that the observer's well-being from an extended alternative in which she is person i for certain equals person i's well-being in that alternative; this is the analogue of Harsanyi's principle of acceptance. Society's well-being, in any simple lottery, is postulated to equal the observer's well-being in the corresponding impartial extended lottery, analogously to Harsanyi's original. Given all this, it is easily shown that society's well-being, in any social alternative, is the average well-being of the individuals in society; so average utilitarianism is true.[8]

Weymark's version of the theorem could be simplified, in that it is not actually necessary to assume that individuals' well-being functions are defined on lotteries, as well as alternatives. (The same is not true of the observer's well-being function, obviously.) Even if individual well-being is only defined for the social alternatives themselves, a utilitarian conclusion can still be derived.[9]

[8] The utilitarian conclusion follows very simply when utility or well-being is taken as primitive in the impartial observer model. The fact that the observer's well-being function is expectational, and satisfies the principle of welfare identity, directly implies that the well-being the observer derives from an impartial extended lottery equals the average individual well-being in the corresponding simple lottery. As Mongin (2001) observes, this is less a 'theorem' than a routine application of decision theory.

[9] A utilitarian conclusion for the social alternatives themselves, that is. To derive a utilitarian conclusion for the lotteries, as well as the alternatives, requires that the individual well-being functions are defined over lotteries (and are expectational).

Weymark argues that his version of the impartial observer theorem, though free from the conceptual problems that plague Harsanyi's original, rests on assumptions that are hard to defend. That well-being is cardinal and interpersonally comparable is controversial enough; but it is the assumption that it is expectational that Weymark finds especially problematic. Why should it be true that the well-being the observer gets from a lottery equals the expected well-being derived from the prizes, i.e. social alternatives? In the context of Harsanyi's original argument this question does not arise, since vNM utility is by definition expectational; but the question is pressing for any non-preference based concept of utility, or well-being.[10]

In section 5, I offer a biological interpretation of Harsanyi's impartial observer theorem, using Weymark's version. It turns out that in a biological setting, the assumptions about well-being or utility that are needed to make the theorem work are readily defensible.

4. BIOLOGICAL PRELIMINARIES

To explain how the veil-of-ignorance concept applies to Mendelian genetics, some biological preliminaries are necessary.

The vast majority of sexually reproducing species are *diploid*, meaning that their cells contain two copies of each chromosome, one paternal and one maternal. The two copies need not be genetically identical; each gene has a number of variants, or *alleles*[11], and the allele at a given slot on the paternal chromosome may differ from the one at the corresponding slot on the maternal chromosome. These chromosomal slots are known as *loci*. If the two alleles at a given locus are the same, the organism is *homozygotic* at that locus, otherwise it is *heterozygotic*. The *genotype* of an organism is a specification of the alleles it contains at one or more loci. So if *A* and *a* are two alleles at a given locus, there are three possible genotypes at that locus: the homozygotes *AA* and *aa*, and the heterozygote *Aa*.

In order to reproduce sexually, diploid organisms produce gametes (e.g. sperm or egg cells) which are haploid, i.e. contain only one of each chromosome pair. Two haploid gametes then fuse to form a diploid zygote, and a new organism is born. Gametes are formed via a cell-division process called *meiosis*, which reduces the chromosome number by half. So only one member of each chromosome pair ends up in a given gamete.

Meiosis proceeds in a number of distinct stages; see Figure 1. Firstly, every chromosome is copied, resulting in a cell that contains two copies

[10] Risse (2002) and Broome (1991) have tried to argue, in a related context, that well-being is 'inherently' expectational, but it is debatable whether their argument succeeds.

[11] In some contexts, biologists use the terms 'gene' and 'allele' more-or-less interchangeably, a practice that is followed here where there is no risk of confusion.

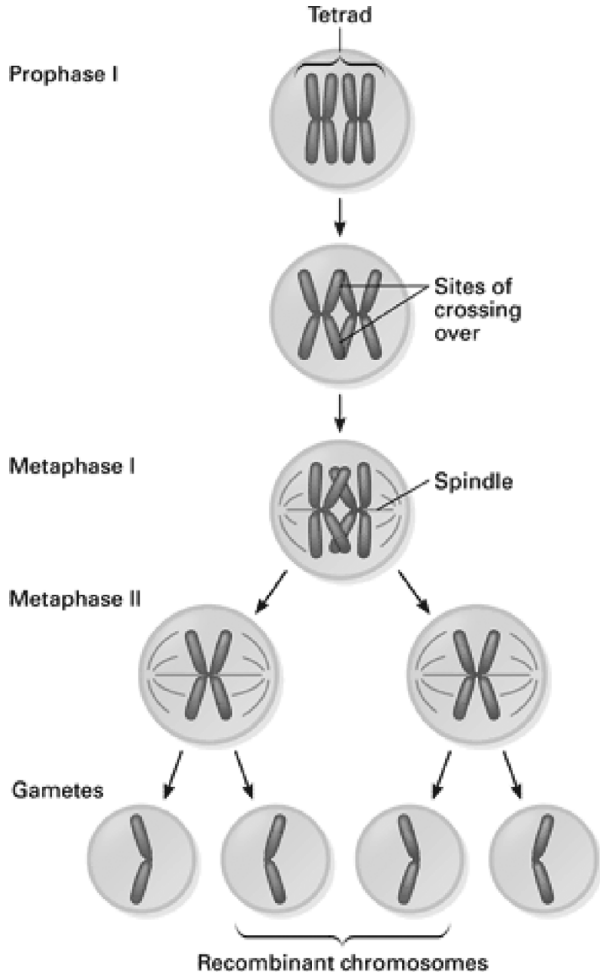

FIGURE 1. Crossing Over

of each maternal and each paternal chromosome. (In Figure 1, which depicts meiosis for a single chromosome pair, this doubling has happened before the diagram starts.) Next a process called 'crossing over' occurs, in which the paternal and maternal chromosomes can swap portions of their genetic material. The paternal and maternal chromosomes in each pair line up alongside each other, break at the same point, exchange parts and then join up. Finally, two rounds of cell division take place, resulting in four haploid gametes each containing one of every chromosome pair. So from a single diploid cell, four haploid gametes are ultimately produced.

Crossing-over means that a chromosome in a gamete will usually not be a perfect copy of either member of the chromosome pair from which it came. This results in a 'shuffling' of genes between parent and daughter chromosomes. If two genes are located at distant ends of a parental chromosome, they may easily be broken up by crossing-over, so will not tend to be transmitted together. But if the genes are located close together on a chromosome, they are likely to be transmitted to a gamete as a unit. Such genes, and the loci they occupy, are said to be 'linked'.

Most of the time meiosis is fair, i.e. each member of a chromosome pair has an equal chance of making it into any gamete. So an *Aa* heterozyote will produce *A* gametes and *a* gametes in equal proportion (on average). This is Mendel's law of segregation, or Mendel's first law. But the law has exceptions. In many species, rogue genes have been discovered which can cheat Mendel's law and get into more than their fair share of gametes. This is known as 'meiotic drive' or 'segregation distortion'; the genes in question are called segregation-distorters. If the *A* allele is a segregation-distorter, then more than half of the gametes produced by the *Aa* heretozygote will be *A* gametes.

Segregation-distorters only have evolutionary consequences when they are in heterozygotes. Suppose there are two alleles *A* and *a* at a locus, and the *A* allele is able to distort segregation in its favour. This means that the ratio of *A* to *a* gametes produced by an *Aa* heterozygote will exceed 1:1, so the *A* allele will increase in frequency in the gene pool, *ceteris paribus*. By contrast an *AA* homozygote only produces *A* gametes anyway (mutation aside), so it makes no difference whether segregation is Mendelian or not.

A segregation-distorter allele enjoys an inherent selective advantage over non-distorters at the same locus. If the *A* allele distorts segregation in its favour and there are no counterbalancing selective pressures, it will sweep to fixation in the population. It is not known how often this has happened in natural populations, since it is very difficult to detect. In other cases the distorter allele does not become fixed, but is maintained in the population at an intermediate frequency. This happens when the distorter allele has negative effects on organismic fitness, which offsets its segregation advantage. Thus suppose that the *A* allele distorts segregation in its favour, but that organisms with the *AA* genotype suffer a fitness disadvantage compared to *Aa* and *aa* organisms. Then for a range of parameter values, the population will evolve to a stable equilibrium in which both alleles are present. When this happens, segregation-distorters can be detected empirically.

Segregation distortion involves a special sort of natural selection, known as 'intra-genomic selection', which takes place between the alleles within a single organism. In the example above, the *A* allele is selected over the *a* allele within *Aa* heterozygotes, since it gets into more than

half the gametes. But this is countered by selection at a higher level, i.e. selection between organisms. *AA* organisms are less fit than *Aa* and *aa*, so their total production of gametes is lower. Thus there are two opposing selective forces at work, operating at different hierarchical levels. In effect, the *A* allele in a heterozygote profits at the expense of its host organism – it reduces the total number of gametes the organism produces, but takes a disproportionate share of the pie for itself.

The above examples assume that a segregation-distorter is a gene at a single locus. This is in principle possible, but empirically most cases of segregation-distortion involve two genes at tightly linked loci, working in concert (cf. Lyttle (1991)). In the fruit fly *Drosophila melanogaster*, an allele at the 'toxin' locus, denoted *Sd*, produces a product that inactivates any gametes that do not produce antidote. Whether a gamete produces antidote depends on which allele it has at a second locus; the *Rsp+* allele does not produce antidote, while the *Rsp* allele does. Since the two loci are tightly linked, the *Sd*/*Rsp* pair constitutes an effective system for meiotic drive. An *Sd*/*Rsp* chromosome will produce toxin that destroys gametes not containing the *Rsp* allele; due to linkage, such gametes will not contain the *Sd* allele either. So both the *Sd* and *Rsp* alleles will achieve greater than fifty percent representation in the successful gametes of their host organism, by destroying gametes containing rival alleles.

5. FAIR MENDELIAN SEGREGATION: EVOLUTIONARY CONSIDERATIONS

Segregation distortion is the exception not the rule; most of the time meiosis is fair. Since fair meiosis involves complex molecular machinery, it seems probable that it is an evolved feature of organisms, so has a Darwinian explanation. This prompts the question: what is the benefit of fair meiosis? Why would natural selection have favoured it?

This question is deceptively simple; we need to ask: 'benefit for whom?' An individual gene clearly benefits if it can distort segregation in its favour – that way it will bequeath more copies to subsequent generations. Imagine again two alleles at a locus, *A* and *a*. Clearly, the *A* allele would prefer a segregation scheme of 4:1 in favour of *A*, for example, over fair 1:1 segregation. Conversely, allele *a* would prefer segregation to be biased in its favour. So neither allele at the locus in question benefits from fair meiosis per se; both would do better if segregation were biased to their advantage.

What about the organism as a whole? Does the whole organism benefit from fair meiosis? The answer is 'yes', in many circumstances. The reason is that a distorter will often have harmful effects on its host organism. Ordinarily a gene which harms its host organism will harm its own reproductive interests, so will be selected against. *But this is not true*

if the gene is a segregation-distorter. A distorter gene may spread by natural selection even if it harms its host organism – which implies a conflict of interest between gene and organism.

To see this point, suppose allele a is initially fixed at a locus, and then allele A arises by mutation. Allele A has two effects: it distorts segregation in its favour, and it reduces the fitness of its host organism. For concreteness, suppose the fitness scheme is: $w(aa) = 10$, $w(Aa) = 9$, $w(AA) = 8$, where these are the numbers of successful gametes produced by an organism of the genotype in question. Suppose that of the 9 successful gametes produced by an Aa heterozygote, 6 are A and 3 are a, on average. Thus the Aa genotype produces fewer successful gametes than the aa genotype, but a disproportionate number of these are A. So despite imposing a fitness cost on its host organism, the mutant A allele is favoured by natural selection when rare. The population evolves to a stable equilibrium where both alleles are present.[12]

This illustrates how segregation distortion can lead to a conflict of interest between a gene and its host organism. Of course a distorter might not harm its host organism; it could be phenotypically neutral, or even beneficial. In these cases the distorter will quickly sweep to fixation in the population, since its segregation advantage is not offset by any negative effects on organismic fitness. But when distorters are maintained in the population in a polymorphic equilibrium, as empirically is often found, then they must be harming their hosts.

The conflict of interest between a distorter allele and its host organism can equally be thought of as a conflict between the distorter allele and the genes at other (unlinked[13]) loci within the organism. If allele A distorts segregation at its own locus, a gene at an unlinked locus is not directly affected. However if the A allele also reduces organismic fitness, as many distorters do, then genes at other loci suffer. They pay the cost of the reduction in fitness but gain no compensating segregation advantage, so end up bequeathing fewer copies to the next generation. Genes at other loci thus benefit if they can somehow prevent the A allele from distorting segregation, so will be under selection to do so.

This is the basis for the explanation of fair meiosis developed by Eshel (1985), who showed that 'modifier' genes at unlinked loci, that have the effect of restoring fair meiosis at the locus undergoing drive, will be

[12] Presuming mating is random, equilibrium is attained when the population-wide frequencies of the A and a alleles are approximately 0.45 and 0.55 respectively. With these frequencies, the marginal fitnesses of the A and a alleles are equal so there is no further evolutionary change.

[13] The qualification 'unlinked' is crucial. At any linked locus, there will be selection for genes which increase allele A's segregation distortion, as they will become preferentially associated with A and will thus gain from the distortion.

favoured by natural selection.[14] So a tug-of-war will ensue: the *A* allele will try to bias segregation in its favour, but alleles at all other unlinked loci in the genome will try to restore meiotic fairness. Since there are many such loci in the genome, they are likely to win the war. The predicted outcome is thus a restoration of fair meiosis. Thus there is a putative evolutionary explanation for why fair meiosis is the rule. This explanation is widely accepted among biologists.

The underlying idea in Eshel's explanation was vividly expressed by Leigh (1977), who spoke of a 'parliament of genes' trying to prevent 'cabals of a few conspiring for their own "selfish profit" at the expense of the "commonwealth"' (p. 4543). Leigh's point was that the bulk of the genes in the genome gain nothing from one of their member distorting segregation at its own locus, and potentially stand to lose a lot, given that distorters often reduce organismic fitness; therefore the genes have a collective interest in enforcing fair meiotic division.

A different (though equivalent) perspective is that fair meiosis acts to equalize the interests of all the genes in the organism. If meiosis is constrained to be fair, then the only way a gene can benefit itself is to boost the fitness (total gametic output) of the whole organism, which benefits all other genes too. By contrast, if a gene can break Mendel's law then it can benefit despite harming its host organism, as we have seen. So fair meiosis acts as a unifying force, preventing internal conflict and leading the organism to behave as a single, cohesive entity.

The fact of meiotic drive, and the evolutionary pressures it gives rise to, remind us that the unity of the individual organism cannot be taken for granted. We naturally regard an individual organism as a cohesive entity, with a unity of purpose, i.e. all its parts work for the common good. This is often justified, but organismic unity is an evolutionary achievement, and is possible only to the extent that meiotic drive (and other forms of intra-genomic conflict) are kept in check (cf. Ridley 2000). An organism in a sexual species is a temporary coalition of genes whose interests do not necessarily overlap, as they are not all transmitted together. Fair meiosis works to align the genes' interests, ensuring they work for the common good. If the parliament of genes could not enforce Mendel's law, harmful genes could spread and organismic integrity would be undermined.

How exactly is fair meiosis enforced? The biochemical details are not well understood, but Haig and Grafen (1991) suggest one possible mechanism. Recall that empirically, most cases of segregation distortion involve a pair of genes at tightly linked loci acting in concert – such as the toxin/antidote *Sd/Rsp* system in *Drosophila* described above. Given this fact, any way of 'unlinking' the two loci, e.g. by increasing the rate

[14] A modifier gene is one which affects the phenotypic expression of some other gene in an organism.

of crossing over, will tend to restore fair meiosis. The *Sd/Rsp* system relies critically on the fact that a gamete containing the *Sd* allele is also likely to contain *Rsp*, and thus to be immune from the toxin. But if the linkage is broken, then the *Sd*'s segregation advantage disappears – for then, the toxin it produces will be as likely to kill a gamete containing a copy of itself as of its rival allele. Destroying linkage thus prevents genes like *Sd* and *Rsp* from forming a selfish cabal at the expense of the rest of the genome. Thus one way to prevent meiotic drive, Haig and Grafen argue, is to destroy linkage, i.e. to force the genes in question to assort independently. Genes on other chromosomes that can achieve this result will be selectively favoured.

More could be said about the evolution of fair meiosis, but the essential points have been made.[15] In a sexual species, there is a potential conflict between any individual gene and its host organism. A gene that distorts segregation in its favour can spread despite reducing its host's fitness. This harms genes at unlinked loci, who are thus under selection to restore fair meiosis if they can. One effect of fair meiosis is to equalize the interests of all the genes, thus ensuring they work for the common good. One way of achieving this is to destroy linkage, given that segregation distortion typically involves linked genes working in concert.

6. MENDELIAN SEGREGATION AND THE VEIL OF IGNORANCE

Finally we are in a position to relate Mendelian genetics to the veil of ignorance. The basis for the analogy is fairly clear. When meiosis is fair, then any given allele does not 'know' whether it will be transmitted to a particular gamete – the chance that it will is fifty percent. So the allele is behind a veil of ignorance with regard to its presence in each gamete, which ensures that its interests are aligned with its host organism, i.e. it will seek to maximize the organism's gametic output. Furthermore, an allele that *is* transmitted to a gamete does not 'know' which other alleles (at unlinked loci) are also in the gamete. By forcing genes behind a veil of ignorance, fair meiosis thus randomizes away the information that alleles would need to profit at the expense of their host, or to form selfish cabals with genes at other loci.

This analogy may seem limited since it says nothing about *choice* from behind the veil, a notion central to the Harsanyi/Rawls argument. But in fact the notion of choice (or preference) is implicit in talk of a gene's 'interests' – which is itself short-hand for talking about what natural selection would favour. This permits the analogy to be elaborated as follows. The alleles in an organism correspond to the individuals

[15] Crow (1991) is a useful review of work on this topic up to 1991; Úbeda and Haig (2005) discuss some recent developments.

in society, in Harsanyi's model. The social alternatives are alternative 'gametic outputs', i.e. specifications of how many successful gametes the organism leaves and which alleles they contain. Each allele has a 'preference order' over the alternatives, determined by how many copies of itself are left in each alternative. The organism itself also has a preference order over the alternatives (analogous to the social preference in Harsanyi's model), determined by the total gametic output, or organismic fitness, in each alternative.

To illustrate, consider a single locus with two alleles A and a. Suppose that for an Aa heterozygote, there are four biologically possible output levels – leaving 0, 1, 2 or 3 successful gametes. All segregation schemes are considered possible. Thus the set of social alternatives S is: {0, A, a, AA, Aa, $a\,A$, aa, AAA, AAa, $Aa\,A$, Aaa, $a\,AA$, $a\,Aa$, $aa\,A$, aaa} where 'AAa', for example, means that the organism leaves three successful gametes, the first two of which contain A and the third a. (Note that the order in which the gametes are produced matters, so AAa and $Aa\,A$ are different alternatives.) The A allele prefers alternative x to y iff it leaves more copies in x than y; thus the A allele prefers alternative AAa to aaa, for example. The a allele has the converse preference. The organism itself is indifferent between AAa and aaa, since its fitness is the same in each (see Table 1).

All alleles at unlinked loci (not modelled here) have the same preference order as the organism; they have no interest per se in whether meiosis is fair at the A/a locus.[16] Note also that the A allele prefers Aa to aaa, despite total gametic output being greater in the latter. This illustrates the fact that the A allele, if permitted to choose, could easily harm the interests of the whole organism.

It bears emphasis that talk of 'interests', 'choice' and 'preference' in this and other evolutionary contexts is fully legitimate, as it can be cashed out precisely and non-metaphorically in terms of natural selection. In saying that the A allele 'prefers' alternative Aa to aaa, we mean that if the A allele could exert causal influence over which of these alternatives obtained, natural selection would lead it to bring about the Aa alternative. Similarly, in saying that the organism is indifferent between AAa and aaa we mean that selection would have no tendency to favour an organism producing one rather than the other of these gametic outputs. A related point is that the 'preference order' of each allele (and the organism) is not primitive, but rather derives from its fitness in each alternative. It is because allele A has a fitness of two in alternative AAa and a fitness

[16] Note that this does not conflict with the standard argument, expounded in section 5, that selection at the organism level (or at unlinked loci) will tend to restore fair meiosis. That argument presumes that the segregation-distorter alleles will have a negative effect on organismic fitness, but this is not being assumed at this juncture.

Rank	A's preference	a's preference	Organism's preference
1.	*AAA*	*aaa*	*AAA*, *AAa*, *Aa A*, *a AA*, *Aaa*, *a Aa*, *aa A*, *aaa*
2.	*AA*, *AAa*, *Aa A*, *a AA*	*aa*, *Aaa*, *a Aa*, *aa A*	*AA*, *Aa*, *a A*, *aa*
3.	*A*, *Aa*, *a A*, *Aaa*, *a Aa*, *aa A*	*a*, *Aa*, *a A*, *AAa*, *Aa A*, *a AA*	*A*, *a*
4.	*aaa*, *aa*, *a*, 0	*AAA*, *AA*, *A*, 0	0

TABLE 1. Preference orders over gametic outputs

of zero in alternative *aaa* that it prefers the former to the latter. So the preference ordering is induced by the fitness function. This is the converse of the usual situation in decision theory, where preferences are primitive and their numerical representations derived. The significance of this will become clear.

Since the *A* and *a* alleles have different preference orders over the social alternatives, and the organism itself has a different preference order again, there is considerable scope for internal conflict – selection will tend to disrupt the integrity of the organism. It is here that fair meiosis comes into play; as we have seen, it has the effect of aligning the interests of all parties.

To represent this in the Harsanyi framework, we first need to introduce lotteries over the set *S*, i.e. probability distributions over possible gametic outputs. This permits two types of randomness or uncertainty to be modelled: about how many successful gametes the organism produces, and about which genes they contain. The former arises because survival and reproduction are stochastic – two organisms of identical genotype won't necessarily enjoy the same reproductive success. Though important in many evolutionary models, this factor is not especially relevant here. The latter arises because of the meiotic process, and is our prime concern. Suppose that meiosis is fair, and that the organism definitely leaves two successful gametes (so there is no uncertainty of the first type.) This picks out a unique lottery over *S*, which has $P(AA) = P(Aa) = P(aA) = P(aa) = \frac{1}{4}$ and zero probability for every other alternative. This is because with fair meiosis, any gamete has an equal chance of receiving the *A* or *a* allele, with independence across gametes. Clearly, any given alternative to fair meiosis, e.g. 3:1 in favour of *A*, also picks out a unique lottery, once the total number of successful gametes has been specified.

What about preference over lotteries? How do we make sense of the idea that allele *A* prefers one lottery to another? The natural way is to assume that an allele evaluates lotteries by the criterion of expected fitness, and so would prefer, i.e. be selected to bring about, the lottery in which its expected fitness is highest; similarly for the organism. Thus allele *A*'s evaluation of the lottery described above, in which two successful gametes are produced and meiosis is fair, equals

$$\frac{1}{4}(2) + \frac{1}{4}(1) + \frac{1}{4}(1) + \frac{1}{4}(0) = 1.$$

Similarly, the organism's evaluation of this lottery equals

$$\frac{1}{4}(2) + \frac{1}{4}(2) + \frac{1}{4}(2) + \frac{1}{4}(2) = 2.$$

Modulo this assumption, it is straightforward to derive a preference order for each allele, and for the organism, over the entire lottery set.

Importantly, there is a real biological rationale for assuming that lotteries are evaluated by expected fitness, namely that this is the evaluation that natural selection will make. When the fitness of an organism (or genotype) is a random quantity, it is well-known that natural selection will select for maximization of expected fitness, given two conditions. These are: (i) that the population is very large, and (ii) the randomness is independent across organisms. Condition (i) is standard in evolutionary theory. To understand condition (ii), suppose an organism of a given genotype leaves 1 or 3 offspring with equi-probability. If each organism of the genotype faces an independent 50:50 gamble on 1 or 3 offspring, e.g. a separate coin is flipped for each, then condition (ii) is satisfied. But if the risk is correlated, e.g. a single coin flip decides whether *all* organisms of the genotype leave 1 or *all* leave 3, then condition (ii) is violated.[17]

In some contexts, assuming independence across organisms (or 'uncorrelated risk') would be unjustified. For example, if an organism's fitness varies randomly because of the weather, the independence assumption would clearly be wrong, since the weather affects many organisms. But in the present context, where our interest is the random variation in an organism's gametic output due to meiosis, the assumption is fully justified. Suppose again that an organism of genotype *Aa* definitely leaves two successful gametes and that meiosis is fair, so $P(AA) = P(Aa) = P(aA) = P(aa) = \frac{1}{4}$. Clearly, this probability distribution is independent across all *Aa* organisms, given how meiosis works. Knowing that one *Aa* organism produced *AA*, for example, tells us nothing about the output of any other. More precisely, if we consider lotteries in which total gametic output is certain, so the uncertainty pertains only to the distribution of the output, then the independence assumption is justified. Expected fitness is thus the right criterion by which an allele, or an organism, should evaluate such lotteries.

We can now use our Harsanyi-style framework to give precise expression to the idea that fair meiosis equalizes the interests of the *A* and *a* alleles, thus preventing conflict. The two alleles have different preference orders over the set S, and thus also over the lottery set ΔS, as we have seen. But consider the subset of lotteries that are meiotically fair, i.e. which are compatible with each allele having an equal chance of entering any gamete. *The A and a alleles will evaluate the meiotically fair lotteries identically, and thus have identical preference orderings over them*. To see this, consider again the lottery in which meiosis is fair, and the organism definitely leaves two successful gametes, i.e. $< P(AA) = P(Aa) = P(aA) = P(aa) = \frac{1}{4} >$. The *A* allele's evaluation of this is $\frac{1}{4}(2) + \frac{1}{4}(1) + \frac{1}{4}(1) + \frac{1}{4}(0) = 1$, which

[17] These two extremes – independence across gametes and perfect correlation – are the opposite ends of a spectrum. Intermediate degrees of correlation are also possible.

is the same as the a allele's evaluation. The same will be true of any meiotically fair lottery.

Note also that the A and a allele's preference ordering over the subset of meiotically fair lotteries *will coincide with that of the organism* – so lotteries will be ranked according to average (or total) gametic output, i.e. in a 'utilitarian' way. To illustrate, compare the meiotically fair lotteries in which the organism definitely leaves one and two successful gametes respectively, i.e. $< P(A) = P(a) = \frac{1}{2} >$ and $< P(AA) = P(Aa) = P(aA) = P(aa) = \frac{1}{4} >$. The organism obviously prefers the latter, as its fitness is twice as high as in the former. Each allele's expected fitness is also twice as high in the latter lottery, so its interests are aligned perfectly with the organism's. Therefore if meiosis is constrained to be fair, the only way that an allele can boost its expected representation in the next generation is by boosting the organism's total gametic output.

This shows that two standard pieces of biological wisdom about fair meiosis are neatly captured by casting the issue in a Harsanyi-style framework. That fair meiosis aligns the interests of competing alleles, with each other and with the host organism, is reflected in the fact that their preference orders coincide over the subset of lotteries that are meiotically fair. The device of fair meiosis, therefore, is the organism's solution to the social contract problem.

This is interesting, and vindicates the idea of treating the alleles in an organism as individuals in a society, with potentially divergent interests. But it does not quite forge a logical link with Harsanyi's own argument. Recall that Harsanyi uses the device of a hypothetical impartial observer to construct a social ordering of *all* lotteries (and thus alternatives) – which turns out to be the utilitarian ordering. We have represented fair meiosis, however, as effecting a restriction on the set of lotteries to the 'meiotically fair' ones; on this subset, each allele's ordering is the utilitarian ordering. There is a similarity here to Harsanyi's argument, but not a exact parallel. Indicative of this is that the notion of an extended lottery, central to Harsanyi's argument, played no role in our representation of fair meiosis. Is it possible to forge a more direct link? It turns out that it is.

7. AN EXPLICIT LINK BETWEEN FAIR MEIOSIS AND THE IMPARTIAL OBSERVER THEOREM

Harsanyi introduces extended lotteries in order to model the notion of impartiality, or justice. He needs to do this since his social alternatives are purely abstract. But in our biological application the alternatives have internal structure – an alternative specifies how many resources (gametes) go to each individual. This permits the notion of impartiality

to be modelled more simply, without the need for extended lotteries, via symmetry considerations.

Call a lottery over gametic outputs *symmetric* if it is invariant under permutation of alleles. Thus for example, the lottery $< P(AAA) = P(aaa) = \frac{1}{2} >$ is symmetric, since if every 'A' is swapped to an 'a' we end up with exactly the same lottery. Note that if a lottery is symmetric, the expected number of A and a alleles it yields is the same, so the two alleles will evaluate it identically.

From an individual allele's point of view, symmetry has an obvious epistemic interpretation. Symmetric lotteries are ones in which the allele is deprived of knowledge about its own identity. For the A allele, the lottery $< P(AAA) = P(aaa) = \frac{1}{2} >$ corresponds to certain knowledge that its host organism will produce three identical successful gametes, but ignorance about who it will be – the lucky or the unlucky one. This interpretation suggests a natural link with Harsanyi's notion of an impartial extended lottery.

Note that all meiotically fair lotteries are symmetric though not vice-versa. Meiotically fair lotteries are ones in which each allele has a given chance of entering any gamete, *with independence across gametes*. So the lottery $< P(AAA) = P(aaa) = \frac{1}{2} >$ is not meiotically fair, since although each allele has an equal chance of being found in (for example) the second successful gamete, independence across gametes is not satisfied – the allele found in the second successful gamete will definitely be found in the first and third. So in meiotically fair lotteries, each allele is deprived of knowledge of its own identity, and also of conditional knowledge of its own identity given any information about the organism's gametic output. So we have: meiotically fair lotteries $\subset$ symmetric lotteries $\subset$ all lotteries.

To each alternative in S, there is exactly one meiotically fair lottery which gives that alternative non-zero probability, and which definitely yields the same total number of gametes as that alternative. Call this the meiotically fair lottery that *corresponds* to the alternative. Thus to the alternative Aa, for example, there corresponds the meiotically fair lottery $< P(AA) = P(Aa) = P(aA) = P(aa) = \frac{1}{4} >$; note that this lottery also corresponds to alternatives AA, aA and aa. More generally, the set S can be partitioned into equivalence classes of alternatives with the same corresponding meiotically fair lotteries. These can be regarded as 'information sets', i.e. alternatives which neither allele can tell apart, under the supposition that meiosis will be fair.

We can then derive an evaluation, and thus an ordering, of the alternatives in S for a given allele 'from behind the meiotic veil'. Ordinarily, the A allele evaluates an alternative by how many copies it leaves in that alternative. But with fair meiosis in place, the allele can't discriminate between alternatives in the same information set; and thus is forced to substitute their evaluation of the corresponding meiotically

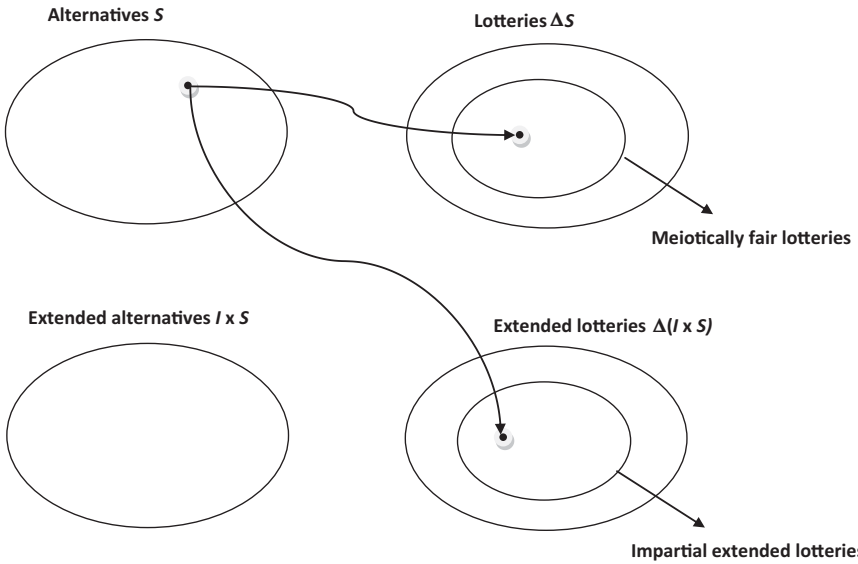

FIGURE 2. Two ways of representing impartiality

fair lottery. Thus the A allele's 'veiled' evaluation of alternative Aa, for example, equals their evaluation of the lottery $< P(AA) = P(Aa) = P(a\,A) = P(aa) = \frac{1}{4} >$, which equals 1; allele a's 'veiled' evaluation is the same. This evaluation then induces a veiled ordering of the alternatives, from best to worst.

Now recall Harsanyi's approach. To each alternative Harsanyi associates a unique impartial extended lottery. Let us apply this to our genetic example, where the alternatives are gametic outputs. Thus for example, Harsanyi's procedure associates to alternative Aa the lottery $< P(Aa, A) = P(Aa, a) = \frac{1}{2} >$; this can be read 'alternative Aa for certain, with equal probability of being person A or person a'. Harsanyi then derives a social ordering of the alternatives, by postulating that it equals the observer's ordering of the corresponding impartial extended lotteries.

An explicit link between the two ways of modelling impartiality is now possible. For each alternative in S, there corresponds to it a unique impartial extended lottery, and a unique meiotically fair lottery (Figure 2). We can then order the alternatives in S two ways: by how the observer would order the corresponding impartial extended lotteries, or by how one of the alleles would order the corresponding meiotically fair lotteries. *But these two give exactly the same result – the average utilitarian ordering.* To see this, compare the two alternatives Aa and aa. The impartial observer, to evaluate these, would compare the lotteries $< P(Aa, A) = P(Aa, a) = \frac{1}{2}$

$>$, and $< P(aa, A) = P(aa, a) = \frac{1}{2} >$. Assuming the principle of welfare identity, and that the observer's valuation function is expectational, the former is evaluated as $\frac{1}{2}(1) + \frac{1}{2}(1) = 1$, the latter as $\frac{1}{2}(0) + \frac{1}{2}(2) = 1$. So the observer deems the alternatives equally good. Now consider how the A allele (or the a allele) would evaluate Aa and aa from behind the meiotic veil. Since both alternatives are in the same information set, the A allele would evaluate them both as $\frac{1}{4}(2) + \frac{1}{4}(1) + \frac{1}{4}(1) + \frac{1}{4}(0) = 1$. So both ways of ordering yield the result that Aa and aa are as good as each other. This result generalizes easily, yielding:

Proposition 1: Let S be the set of alternative gametic outputs. Let R be the ordering of S given by the impartial observer's ordering of the corresponding impartial alternative lotteries (assuming the principle of welfare identity, and that the observer's valuation function is expectational). Let R' be the ordering of S given by any allele's ordering of the corresponding meiotically fair lottery. Then, $R = R'$.

Proof: see Appendix

Proposition 1 shows that the link between Harsanyi's way of modelling social justice, via the device of an impartial observer, and the natural way of modelling genomic justice, via a restriction on the set of lotteries to the meiotically fair ones, is actually quite intimate. In reality, the two amount to the same thing: they capture the notion of impartiality in different, but equivalent, ways. The main difference is the greater generality of Harsanyi's approach, since it does not require that the social alternatives have structure.

A second, minor difference is that Harsanyi's way of deriving a social ordering of the alternatives derives from a more general ordering of the lotteries; while the 'meiotically fair' lottery approach yields a direct ordering of the alternatives. But this difference is more real than apparent. On the meiotically fair approach, it would be straightforward to extend the 'social' ordering to the full lottery set ΔS, if we wished, simply by stipulating that a lottery's valuation is its expected valuation.

Importantly, proposition 1 could be strengthened by considering symmetric lotteries rather than meiotically fair lotteries. To each alternative in S, there corresponds a set of symmetric lotteries each of which definitely yields the same total number of gametes as that alternative, but only one of which is meiotically fair. However, an allele will evaluate all of these symmetric lotteries identically, by the average utilitarian rule. Therefore, proposition 1 would be true if for 'meiotically fair' lottery we substituted any member of the class of symmetric lotteries that correspond to the alternative in question. That is, restricting the permissible lotteries to the meiotically fair ones is a way, but not the

only way, of forcing alleles to substitute the organism's interest for their own.

This last observation raises an important question. We have been assuming, with biological orthodoxy, that the function of fair meiosis is to unify the interests of the genes in an organism. But fair meiosis, which amounts to flipping a separate fair coin for each gamete that is produced, is not the only way to achieve this task. Instead, an *Aa* heterozygote could flip one coin for all the gametes to be produced, i.e. they all get allele A or all get allele a, with equal probability. If two successful gametes are definitely produced, the resulting lottery is $< P(AA) = P(aa) = \frac{1}{2} >$ – which is symmetric though not meiotically fair. A veil of ignorance of this type would also lead alleles to evaluate lotteries by average gametic output. Why did evolution not solve the problem of conflicting interests this way?

I do not know the answer to this question. It may be that using a separate coin flip for each gamete is the simplest way to implement symmetry, or there may be an underlying evolutionary reason why the allocation of genes to gametes should be independent across gametes, i.e. some adaptive advantage to this independence. Whatever the answer, this question highlights the fact that not all aspects of diploid genetics can be accounted for by the parallel between Harsanyi's veil of ignorance and its Mendelian counterpart. This is not really surprising; if anything it is surprising that the parallel extends as far as it does.

What exactly is the upshot of proposition 1? It shows, in a precise manner, that the way fair meiosis leads alleles to align their interest with the whole organism is identical to the way that the veil of ignorance leads Harsanyi's impartial observer to align her interests with those of society as a whole – where the latter are defined by the average utilitarian rule. Of course the veil of ignorance is a mere thought experiment for Harsanyi and Rawls, so it is remarkable that the underlying principle behind it – that randomization can serve to align the interests of competing agents – finds an embodiment in real biological systems. Note also that the biological instantiation of the principle corresponds better to Harsanyi's than to Rawls's version of the argument, in that the former's utilitarian conclusion clearly holds in the biological case. From behind the meiotic veil, an allele will use the average utilitarian rule to evaluate alternatives, given standard biological assumptions.

In a sense, our biological version of the impartial observer theorem is actually superior to Harsanyi's original. Recall the Sen/Weymark challenge to Harsanyi: to justify properly his assumption that utility (well-being) is both expectational and inter-personally comparable. This challenge is straightforward to meet in the biological case, where utility is replaced by fitness, and the genes in the organism are the individuals in society. Fitness is trivially (fully) comparable across genes (and

organisms); indeed the whole point of the fitness concept is to permit comparisons between biological units (genes, genotypes and organisms). A gene's fitness, as understood here, is simply the number of copies left in the next generation, and an organism's fitness means its total gametic contribution to the next generation. So inter-personal comparability is unproblematic in the biological case.

Moreover, the assumption that the valuation function is expectational, which is crucial to Harsanyi's theorem, has a real biological rationale, as noted earlier. For when risk is uncorrelated, the 'right' way for a gene (or an organism) to evaluate a lottery is to use the expected value of that lottery – in that this is the evaluation that matters to natural selection. With uncorrelated risk, a gene which codes for lottery L_1 will be selected over one which codes for L_2 if and only if the expected number of copies it leaves in L_1 is greater than in L_2. So the 'expectational' nature of the valuation function, for which Harsanyi offered no good argument, admits of a genuine rationale in our biological application of the theorem.

There is a certain irony in this. At first blush, the idea of applying Harsanyi's theorem to Mendelian genetics may seem implausible, a rather strained analogy. But closer examination reveals that the connection is quite intimate – the principle that Harsanyi discovered really is instantiated in genetic systems. Moreover in the genetic case, the Sen/Weymark criticism cuts no ice. Sen and Weymark were no doubt right that Harsanyi's theorem does not amount to a 'proof' of classical utilitarianism, but Harsanyi's underlying point – that the veil of ignorance leads to a utilitarian evaluation rule given certain assumptions – was of course correct. The problem was that the assumptions in question were ones that Harsanyi could not justify. But in the biological case the corresponding assumptions can be justified, and the analogue of utilitarianism – that organisms, and their constituent genes, should try to maximize total gametic output – actually holds true.

8. COMPARISON WITH RIDLEY'S ANALYSIS

In his book *Mendel's Demon*, the biologist Mark Ridley (2000) provides an extended discussion of how the veil-of-ignorance concept applies to genetics. My analysis differs from Ridley's in two main ways. Firstly, Ridley focuses exclusively on Rawls's rather than Harsanyi's version of the veil-of-ignorance argument. However Harsanyi's version is more relevant for the parallel with genetics, both because it is formally elaborated, which allows the parallel to be made precise, and because his decision-theoretic assumptions make good sense in a biological context. Indicative of this is that Harsanyi's utilitarian conclusion holds true in biological systems with fair meiosis, while Rawls's maximin conclusion does not.

Secondly, Ridley's view of how the veil-of-ignorance concept applies to genetics is different from our own. Recall from section 4 that two sorts of randomization occur in genetics. Firstly, one of each chromosome pair is allocated at random to a gamete, when meiosis is fair. Secondly, crossing over breaks up linkage, which means that whether an allele at one chromosomal locus gets in to a particular gamete is independent of whether an allele at another locus does. As a result, selfish cabals that subvert the group's interests cannot form. This helps keep meiosis fair, given that empirically, successful distortion of segregation requires genes at two loci to work in concert (as in the *Sd/Rsp* system in *Drosophila melanogaster*). So the second sort of randomization helps stabilize the first.

In our analysis, it is the first sort of randomization that occupies centre-stage, for it is here that the link with the Harsanyi/Rawls argument is strongest. Just as the veil of ignorance leads Harsanyi's impartial observer to choose the option that maximizes society's total welfare, so the Mendelian veil of ignorance (i.e. fair meiosis) leads genes to choose the option that maximizes their organism's total gametic output, thus equalizing their individual interests with that of the collective.

However, Ridley's emphasis is on the second sort of randomization, i.e. the genetic recombination caused by crossing over. It is here that he thinks the analogy with Rawls's argument works best (p. 200). But this seems questionable. The main effect of recombination is to deprive genes of information about the identity of genes at other loci, which prevents selfish cabals forming, as Ridley emphasizes. This is an important point, but it has no clear counterpart in the Rawls/Harsanyi argument. The latter involves a single individual – the impartial observer – uncertain about which member of society he will become. The uncertainty does not concern the characteristics of *other* members of society – or at least, any such uncertainty is strictly irrelevant to the decision problem that the impartial observer faces. So the second sort of randomization, at least in so far as its cabal-stopping consequences are what matters, has no parallel in the Harsanyi/Rawls story.[18]

My disagreement with Ridley over how the veil-of-ignorance concept applies to genetics is related to the difference between proximate and ultimate explanations. Ridley emphasizes that recombination prevents

[18] To be fair to Ridley, he also discusses another consequence of crossing over, to which this criticism does not apply. Following Haig and Grafen (1991), he argues that meoisis is designed to prevent the spread of 'sister killer' genes, which gain a transmission advantage by causing their host gamete to kill its sister gamete after meiotic cell division. Crossing over frustrates such sister killer genes, as it means that a putative sister killer doesn't 'know' whether a copy of itself will be found in the sister gamete or not, so is just as likely to harm as to help itself. This consequence of crossing over is distinct from its cabal-stopping consequences. Thanks to an anonymous referee for this observation.

selfish cabals forming, and thus deprives genes of the information they would need to cheat Mendel's first law. This helps keep meiosis fair, given empirical facts about how systems of segregation-distortion actually work. So the second sort of randomization is part of the proximate mechanism by which fair meiosis is maintained. But this says nothing about why fair meiosis is adaptively advantageous in the first place. The first sort of randomization, by contrast, addresses this ultimate question. By equalizing the chances that each allele at a locus will enter a given gamete, the interests of genes are aligned, with each other and with the whole organism. Selection at the organism level will therefore tend to produce fair meiosis.

The contrast between the two types of randomization, therefore, is in part a contrast between a particular mechanism by which fair meiosis is in fact stabilized, and the evolutionary consequences that follow from meiosis being fair, however this fairness is achieved. It makes sense that a biological version of the Harsanyi/Rawls argument should have an ultimate orientation. For the Harsanyi/Rawls argument involves the notion of utility, or welfare, applied to both individuals and whole societies. The biological analogue is the notion of fitness, which also applies to both individual genes and whole organisms. But the notion of fitness is the paradigmatically ultimate notion; questions of fitness concern evolutionary consequences, not proximate mechanisms.

Despite these criticisms, Ridley's treatment is insightful and provided the inspiration for the foregoing analysis. Ridley concludes his discussion by saying that 'Rawls's mechanism in a way applies more powerfully to genes than it does to human beings' (p. 200). If 'Rawls' is replaced with 'Harsanyi', then I think this conclusion is exactly right. As we have seen, the impartial observer argument as applied to humans is in fact fraught with difficulty. The assumptions that Harsanyi uses in his theorem are difficult to justify in the rational choice context that he was operating in. In the biological context, by contrast, the assumptions are straightforward to justify, and the utilitarian conclusion actually holds true. Individual genes do in fact act to maximize their organism's gametic output, as a result of the fairness of meiosis, so they behave like utilitarian agents.

9. CONCLUSION

This paper has explored a parallel between the veil-of-ignorance concept in social ethics and in evolutionary genetics. I have argued that the parallel is a genuine one that runs deep and has real explanatory power. In particular, there is an intriguing biological analogue of Harsanyi's impartial observer argument that is free from the difficulties that plague Harsanyi's original. The principle that Harsanyi discovered is actually instantiated in the genetic systems of sexually reproducing organisms,

and ensures that their constituent genes work for the good of the whole organism.

The parallel I have developed derives from the fact that utility and fitness play isomorphic roles in rational choice theory and evolutionary theory respectively. This role-isomorphism has been noted before by many authors, particularly in relation to decision-making in strategic contexts, but has only rarely been applied in relation to social choice.[19] However there is no good reason for this, since the basic premise of social choice – the existence of individuals in a society with divergent interests – is directly applicable to many biological systems. Future work will be needed to tell whether this conceptual link can be fruitfully exploited.

REFERENCES

Arrow, K. 1951. *Social Choice and Individual Values*. New York: Wiley.

Binmore, K. 2006. *Natural Justice*. Oxford: Oxford University Press.

Broome, J. 1991. *Weighing Goods*. Oxford: Blackwell.

Conradt, L. and C. List 2009. Group decisions in humans and animals: a survey. *Philosophical Transactions of the Royal Society B* 364: 719–742.

Crow, J.F. 1991. Why is Mendelian segregation so exact? *Bioessays* 13: 305–312.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Eshel, I. 1985. Evolutionary genetic stability of Mendelian segregation and the role of free recombination in the chromosomal system. *American Naturalist* 125: 412–420.

Frank, S.A. 2003. Repression of competition and the evolution of cooperation. *Evolution* 57: 693–705.

Godfrey-Smith, P. 2009. *Darwinian Populations and Natural Selection*. Oxford: Oxford University Press.

Haig, D. 1997. The social gene. In *Behavioural Ecology*, 4th edition, eds. J.R. Krebs and N.B. Davies, 284–304. Oxford: Blackwell.

Haig, D. and C.T. Bergstrom 1995. Multiple mating, sperm competition and meiotic drive. *Journal of Evolutionary Biology* 8: 265–282.

Haig, D. and A. Grafen 1991. Genetic scrambling as a defense against meiotic drive. *Journal of Theoretical Biology* 153: 531–558.

Hamilton, W.D. 1964. The genetical evolution of social behaviour, i and ii. *Journal of Theoretical Biology* 7: 1–52.

Harsanyi, J.C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434–435.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Leigh, E.G. jr. 1971. *Adaptation and Diversity*. San Francisco: Freeman Cooper.

Leigh, E.G. jr. 1977. How does selection reconcile individual advantage with the good of the group? *Proceedings of the National Academy of the Sciences USA* 74: 4542–4546.

List, C. 2004. Democracy in animal groups: a political science perspective. *Trends in Ecology and Evolution* 19: 168–9.

Lyttle, T.W. 1991. Segregation distorters. *Annual Review of Genetics* 25: 511–557.

[19] Okasha (2009) explores conceptual connections between social choice theory and the theory of multi-level selection, which deals with natural selection in a hierarchical world. Other applications of social choice-theoretic ideas to the biological realm include Conradt and List (2009) and List (2004), which focus on group decision making.

Mongin, P. 2001. The impartial observer theorem of social ethics. *Economics and Philosophy* 17: 147–149.
Okasha, S. 2009. Individuals, groups, fitness and utility: multi-level selection meets social choice theory. *Biology and Philosophy* 24: 561–584.
Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Ridley, M. 2000. *Mendel's Demon: Gene Justice and the Complexity of Life*. London: Weidenfeld and Nicholson.
Risse, M. 2002. Harsanyi's 'utilitarian theorem' and utilitarianism. *Noûs* 36: 550–577.
Roemer, J.E. 1998. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.
Sen, A.K. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7: 243–262.
Sen, A.K. 1977. Non-linear social welfare functions: a reply to Professor Harsanyi. In *Foundational Problems in the Special Sciences*, eds. R. Butts and J. Hintikka, 297–302. Dordrecht: Reidel.
Sen, A.K. 1986. Social choice theory. In *Handbook of Mathematical Economics III*, eds. M.D. Intriligator and K.J. Arrow, 1073–1181. Amsterdam: North Holland.
Skyrms, B. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Sober, E. 1998. Three differences between evolution and deliberation. In *Modelling Rationality, Morality and Evolution*, ed. P. Danielson, 408–422. Oxford: Oxford University Press.
Úbeda, F. and D. Haig 2005. On the evolutionary stability of Mendelian Segregation. *Genetics* 170: 1345–1357.
Vickrey, W.S. 1945. Measuring marginal utility by reaction to risk. *Econometrica* 13: 319–333.
Weymark, J. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-being*, eds. J. Elster and J.E. Roemer, 225–320. Cambridge: Cambridge University Press.

A. PROOF OF PROPOSITION 1

S is the set of alternative gametic outputs. So each $x \in S$ is a finite sequence of As and as, e.g. $Aa\,Aa\,a\,AAa$. Let N_x be the total number of gametes produced in alternative x. Let A_x be the total number of A gametes produced in alternative x. Let x' be the impartial extended lottery that corresponds to x. So $x' \equiv < P(x, A) = P(x, a) = \frac{1}{2} >$. Let x'' be the meiotically fair lottery that corresponds to x. Let V_o be the impartial observer's valuation function. Let V_A be the A allele's valuation function. Let V_a be the a allele's valuation function. We wish to show that $V_o(x') = V_A(x'')$.

Consider the impartial extended lottery $x' \equiv < P(x, A) = P(x, a) = \frac{1}{2} >$. By the principle of welfare identity, and the fact that $V_o(x')$ is expectational, we have: $V_o(x') = \frac{1}{2}V_A(x) + \frac{1}{2}V_a(x)$. But $V_A(x) = A_x$ and $V_a(x) = N_x - A_x$. Therefore $V_o(x') = \frac{1}{2}N_x$.

Next, consider the meiotically fair lottery x''. x'' is an equiprobable lottery over 2^{N_x} alternatives in S. In each of these alternatives, the number of A alleles ranges from 0 to N_x. The proportion of alternatives in which there are exactly z A alleles is: $\frac{1}{2^{N_x}}\binom{N_x}{z}$. So $V_A(x'') = \frac{1}{2^{N_x}} \cdot \sum_{z=0}^{N_x} z.\binom{N_x}{z} = \frac{1}{2}N_x$. Therefore $V_o(x') = V_A(x'') = \frac{1}{2}N_x$.

That is, for any alternative $x \in S$, the impartial observer's valuation of the corresponding impartial extended lottery x' equals the A allele's valuation of the corresponding meiotically fair lottery x''.

QED